

# ICR Project Summary

SIG	Project Name	Project Summary	Name	Role	Center
Data Analysis and Statistical Tools	Distance-Weighted Discrimination	DWD is a tool that performs statistical corrections to reduce systematic biases that are due to different sources of RNA, different batches of microarrays and especially different microarray platforms. This tool, currently implemented in MatLab, will be ported to an open source, caBIG-compliant platform this year.	Steve Marron	Developer	Lineberger
			Everett Zhou	Developer	Lineberger
			Michael Nebozhyn	Adopter	Wistar
	GenePattern	GenePattern is a flexible analysis platform developed to support multidisciplinary genomic research programs. Its architecture and environment are expressly designed to allow rapid prototyping and integration of new technologies. For caBIG, an adapter to GenePattern will be created to allow the application to read data directly from caBIO/caArray.	Michael Showe	Adopter	Wistar
			Ted Liefeld	Developer	MIT/Broad
			Michael Reich	Developer	MIT/Broad
			Jomol Mathew	Adopter	New York
	Magellan	Magellan is a web-based system that allows biologists to perform complex analyses on heterogeneous data in an environment that does not require a background in computer programming or statistics. Stored data and annotations are treated as abstract entities such that arbitrary, user-defined types of information can be stored. In the context of caBIG, Magellan will consume of data from caBIO/caArray and adapt its architecture to Silver-level compliance.	Judith Goldberg	Adopter	New York
			David Fenstermacher	Adopter	Penn
			Craig Street	Adopter	Penn
			Chris Kingsley	Developer	UC San Francisco
	VISDA	Visual and Statistical Data Analyzer (VISDA) has been developed with the goal of revealing all of the interesting patterns in a dataset. The main application of VISDA is for multivariate cluster modeling, discovery, and visualization, particularly for data sets living in high dimensional space. Currently implemented in MatLab, VISDA will be ported to a caBIG-compliant, open-source architecture.	Ajay Jain	Developer	UC San Francisco
			Joseph Wang	Developer	Georgetown
			Malik Yousef	Adopter	Wistar
			Michael Nebozhyn	Adopter	Wistar
Genome Annotation	Cancer Molecular Pages	This system will be based on technology originally developed by the Joint Center for Structural Genomics, to aid cancer researchers in with keeping up with all of the information being generated on a gene or gene set of interest. It is a fully functional database and automated annotation system combining automated computer-based annotations and automated data collection from experimental stations and set of web based visualization tools. This system will be ported to a caBIG-compliant architecture this year and will consume of data from caArray.	Michael Showe	Adopter	Wistar
			Kutbuddin Doctor	Developer	Burnham
	FunctionExpress	FunctionExpress is an environment for the integrated analysis and visualization of complementary data sets. The system provides a mechanism for regular updates of integrated annotation data that can be readily associated with probes on microarrays. The application allows for plotting of raw and transformed microarray data, for viewing orthologous probesets from other experiments, and for creating literature-based gene networks. This system, currently implemented in a 2-tier, proprietary archicture will be ported to an open source, caBIG-compliant architecture this year.	TBD	Adopter	TBD
			Rakesh Nagarajan	Developer	Wash U
			John Rux	Adopter	Wistar
	GOMiner	GOMiner leverages the Gene Ontology (GO) to identify the biological processes, functions and components represented in gene lists. Instead of analyzing microarray results with a gene-by-gene approach, GoMiner classifies the genes into biologically coherent categories and assesses these categories. For caBIG, GoMiner will be adapted to a caBIG-compliant architecture and made available as a web service.	Harold Riethman	Adopter	Wistar
			David Kane	Developer	NCI - CCR
			John Rux	Adopter	Wistar
	Human PromoterDB	The HapMap database is a repository of human SNPs, their genotypes, and the linkage disequilibrium relationships among them. The Vertebrate Promoter Database (VPD) is a curated resource for vertebrate transcription factor binding sites and their corresponding regulatory regions.	Harold Riethman	Adopter	Wistar
			Lincoln Stein	Developer	Cold Spring
			Alex Lash (PromoterDB)	Adopter	Sloan

## ICR Project Summary

	<b>HapMap, PromoterDB</b>	Both of these datasets will be made available through caBIO in this caBIG project.	Harold Riethman (HapMap)	Adopter	Wistar
	<b>PIR</b>	The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research. PIR maintains the Protein Sequence Database (PSD), an annotated protein database containing over 283,000 sequences covering the entire taxonomic range. PIR is also a member of the UniProt consortium, the central international resource of protein sequence and function that unifies the PIR, Swiss-Prot, and TrEMBL databases. For caBIG, the PIR database will be grid-enabled to demonstrate how such a datasource can be discovered and consumed in a grid environment.	Cathy Wu	Developer	Georgetown
			Craig Street	Adopter	Penn
			David Fenstermacher	Adopter	Penn
	<b>SEED</b>	SEED is a framework that supports peer-to-peer annotation of genomes. Investigators can work independently on their own instances of the SEED database and synchronize their work when desired or update code versions quickly via the network. Several hundred microbial organisms are in SEED now and pipelines for high-throughput processing, e.g., BLASTing, exist for rapidly including new organisms. For caBIG, access to curated SEED data will be made available through a caBIG-compliant interface. In addition, eukaryotic data will be added to SEED.	Cathy Wu	Adopter	Georgetown
			Ed Frank	Developer	Holden
<b>Microarray Repositories</b>	<b>caArray</b>	NCICB is developing their next generation microarray data repository, termed caArray. Phase I of the system, to be released in September, will be caBIG-compliant and will contain the following key features: MIAME 1.1 compliant; support for MAGE-ML import and export; utilities for the submission and retrieval of Affymetrix and GenePix native file formats; use controlled vocabularies; accessibility through a MAGE-OM API. Many Centers in the ICR Workspace are interested in adopting caArray.	Arnie Miles	Adopter	Georgetown
			Michael Showe	Adopter	Wistar
			Walter Mankowitz	Adopter	Wistar
			Wen Hwai Horng	Adopter	Wistar
			Jomol Mathew	Adopter	New York
			Judith Goldberg	Adopter	New York
<b>Microarray Repositories</b>	<b>NCI-60 Data Sharing</b>	The NCI-60 are 60 diverse human cancer cell lines used by the NCI Developmental Therapeutics Program to screen >100,000 chemical compounds since 1990 for anticancer activity. The Weinstein and other laboratories have assessed potential molecular targets and modulators of activity in those cells one or a few at a time and have developed a number of "omics" databases for this data. For this project, these datasets will be made available through standards and interfaces developed/approved by caBIG.	David Kane	Developer	CCR
			Alex Lash	Adopter	Sloan
	<b>Zebrafish microarray data sharing</b>	The Thomas Jefferson University Zebrafish Microarray Service provides microarray data generation services for the community of scientists using zebrafish as a model organism for the study of cancer. They have developed a custom microarray using the commercial Zebrafish oligo library (Compugen/Sigma-Genosys). Through this caBIG project, Thomas Jefferson will make datasets from this repository available to the research community by identifying sharable datasets, creating a web application to select these datasets, and making these datasets available via a caBIG-defined exchange format.	Alex Lash	Adopter	Sloan
			Jack London	Developer	Thomas Jefferson
	<b>Pathways Tool Development</b>	Sloan has developed resources aimed at aiding life science researchers in visualizing and interacting with information in the context of biological pathways. These resources are BioPAX, a common exchange format for pathways data; cPath, a database focused on protein-protein interactions, and Cytoscape, a bioinformatics software platform for visualizing molecular interaction networks and integrating these interactions with gene expression profiles and other stat data. Sloan will leverage these existing resources to create a set of tools that will allow users to view pathway data in the context of caBIG annotation and expression data.	Shannon McWeeney	Adopter	Oregon Health
			Gary Bader	Developer	Sloan
		QPACA stands for Quantitative Pathway Analysis in Cancer. It is a pathway modeling and analysis	Shannon McWeeney	Adopter	Oregon Health

## ICR Project Summary

Pathways	QPACA	system that supports exploration of quantitative biological data in the context of a pathway description. At the center of the system is a pathway representation that enables visualization and computational analysis of pathway structure. For caBIG, QPACA will be made interoperable with Magellan, a web-based system that allows the biologist to perform complex analysis on heterogeneous data and that is being made interoperable with caBIO and caArray. QPACA will also be made interoperable with the pathway exchange standard that is being defined by caBIG.	Ajay Jain	Developer	UC San Francisco
	Reactome (GKB) Data	The Reactome database (formerly known as GKB), developed at Cold Spring Harbor, is a curated database of fundamental biological pathways in human which uses strict rules of assertion and evidence tracking to ensure a consistent high quality product. For caBIG, this data will be made available through a pathways data exchange format that is defined by caBIG.	Lincoln Stein	Developer	Cold Spring
			Brian Gilman	Developer	Cold Spring/Panther Informatics
			Alex Lash	Adopter	Sloan
Proteomics	Proteomics LIMS	This project will be focused on the creation of a caBIG-compliant proteomics Laboratory Information Management System (LIMS). The initial version will track the lab processes relevant to 2D gel electrophoresis but the schema will support the addition of new data types as they emerge. The availability of an open source proteomics LIMS will ultimately allow a distributed development model in which additional modules can be contributed by other centers.	Michael Ochs	Developer	Fox Chase
			Tom Moloshok	Developer	Fox Chase
			Steve Eschrich	Adopter	Moffitt
	Q5	The Q5 algorithm supports probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. The algorithm employs Principal Components Analysis (PCA) followed by Linear Discriminant Analysis (LDA) on whole spectrum Surface-Enhanced Laser Desorption/Ionization Time of Flight (SELDI-TOF) Mass Spectrometry (MS) data. Currently implemented in MatLab, Q5 will be ported to a caBIG-compliant, open-source architecture.	David Jewell	Developer	Dartmouth
			Shannon McWeeney	Adopter	Oregon Health
	RProteomics	In this project, R libraries will be developed to post-process MALDI-TOF and SELDI-TOF data. These libraries will be incorporated into a caBIG-compliant, user-friendly system that will aid cancer researchers in the processing of their TOF data. RProteomics will be developed as a grid reference implementation to demonstrate workflow in a grid environment.	Patrick McConnell	Developer	Duke
			Shannon McWeeney	Adopter	Oregon Health
			Craig Street	Adopter	Penn
Translational	TrAPSS	TrAPSS is a system comprised of several tools that aid scientists who are searching for the genetic mutation or mutations that cause a defect or disease. The system offers support for almost all areas of a mutation discovery project from the creation and prioritization of a large candidate gene list, to the selection, ordering, and managing of primer pairs, and even support for SSCP assay results.	David Fenstermacher	Adopter	Penn
			John Rux	Adopter	Wistar
			Harold Riethman	Adopter	Wistar
			Terry Braun	Developer	U Iowa Holden